

Protein Motif Searches and Regular Expressions

Exercise 6

6.1 Using InterPro domain searches to identify unannotated kinesin motor proteins.

For this exercise use <http://fungidb.org>

a. Identify all genes annotated as hypothetical in *Phytophthora infestans*.

Hint: use the full text search and look for genes with the word “hypothetical” in their product names.

Identify Genes based on Text (product name, notes, etc.)

Organism [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

- Agaricomycetes
- Blastocladiomycetes
- Chytridiomycetes
- Eurotiomycetes
- Leotiomycetes
- Oomycetes
 - Hyaloperonospora
 - Phytophthora
 - Phytophthora capsici
 - Phytophthora cinnamomi
 - Phytophthora infestans
 - Phytophthora infestans T30-4
 - Phytophthora parasitica
 - Phytophthora ramorum
 - Phytophthora sojae
 - Pythium
 - Saprolegnia
- Pneumocystidomycetes
- Pucciniomycetes
- Saccharomycetes
- Schizosaccharomycetes
- Sordariomycetes
- Tremellomycetes
- Ustilaginomycetes
- Zygomycetes

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

Text term (use * as wildcard)

Fields

- Alias
- EC descriptions
- Gene ID
- Gene notes
- Gene product
- GO terms and definitions
- Metabolic pathway names and descriptions
- Protein domain names and descriptions
- Similar proteins (BLAST hits v. NRDB/PDB)
- User comments

[select all](#) | [clear all](#)

Advanced Parameters

Get Answer

b. How many of these hypothetical genes have a kinesin-motor protein InterPro domain?

Hint: add a step to the strategy. Go to the “Interpro Domain” search under similarity/pattern, start typing the work kinesin and it should autocomplete.

Add Step

- Run a new Search for
- Transform by Orthology
- Add contents of Basket
- Add existing Strategy
- Filter by assigned Weight

Genes

- Genes
- Genomic Segments (DNA Motif)
- SNPs
- ORFs
- SAGE Tags

Text, IDs, Species

- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution
- Population Biology

Protein Motif Pattern

- Interpro Domain
- BLAST

Add Step 2 : InterPro Domain

Organism: select all | clear all | expand all | collapse all | reset to default

- Agaricomycetes
- Basidiomycetes
- Chytridiomycetes
- Eurotiomycetes
- Lecanidomycetes
- Oomycetes
- Phylogenomycetes
- Phytophthora
- Phytophthora capsici
- Phytophthora cinnamomi
- Phytophthora infestans
- Phytophthora parasitica
- Phytophthora ramorum
- Phytophthora sojae
- Pythium
- Saprolegnia
- Phaeocystidomycetes
- Pucciniomycetes
- Saccharomycetes
- Sordariomycetes
- Sordariomycetes
- Tremellomycetes
- Ustilaginomycetes
- Zygomycetes

Domain Database: INTERPRO

Specific Domain(s): IPR001752: Kinesin_motor_dom

Combine Genes in Step 1 with Genes in Step 2:

- 1 Intersect 2
- 1 Union 2
- 1 Minus 2
- 2 Minus 1
- Relative to 2, using genomic colocation

9 Genes from Step 2 [Add 9 Genes to Basket](#) | [Download 9 Genes](#)

Strategy: Text

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	Ajellomyces		Aspergillus										Batrachochytrium	Botryotinia	Candida	
		<i>A. capsulatus</i> (nr Genes: 0)	<i>A. macrogynus</i>	<i>A. aculeatus</i>	<i>A. carbonarius</i>	<i>A. clavatus</i>	<i>A. flavus</i>	<i>A. fumigatus</i>	<i>A. nidulans</i>	<i>A. niger</i> (nr Genes: 0)	<i>A. terreus</i>	<i>B. dendrobatidis</i>	<i>B. fuckeliana</i>	<i>C. albicans</i>	<i>C. glabrata</i>		
9	9	G186AR	NAm1	ATCC 38327	ATCC 16872	ITEM 5010	NRRL 1	NRRL3357	Af293	FGSC A4	ATCC 1015	CBS 513.88	NIH2624	JEL423	B05.10	SC5314	CBS 138
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

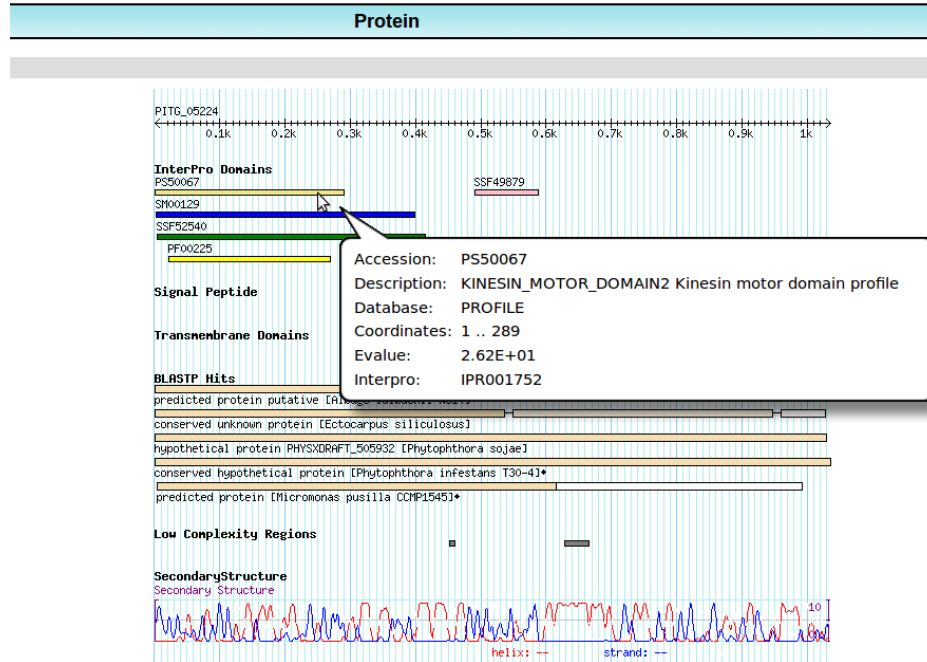
Gene Results | Genome View

Advanced Paging Add Columns

Gene ID	Genomic Location	Product Description
PTG_00478	PinFT30-4_SC0001: 2,501,858 - 2,505,298 (-)	conserved hypothetical protein
PTG_02538	PinFT30-4_SC0003: 2,468,713 - 2,469,267 (+)	conserved hypothetical protein
PTG_05224	PinFT30-4_SC0007: 1,596,587 - 1,600,159 (-)	conserved hypothetical protein
PTG_10158	PinFT30-4_SC0017: 2,018,080 - 2,021,529 (+)	conserved hypothetical protein
PTG_10474	PinFT30-4_SC0018: 1,467,527 - 1,468,650 (+)	hypothetical protein
PTG_11652	PinFT30-4_SC0022: 1,042,072 - 1,044,886 (+)	conserved hypothetical protein
PTG_16513	PinFT30-4_SC0044: 896,186 - 897,436 (+)	conserved hypothetical protein
PTG_19553	PinFT30-4_SC0098: 255,381 - 256,027 (+)	conserved hypothetical protein
PTG_21476	PinFT30-4_SC0374: 18,707 - 19,392 (+)	conserved hypothetical protein

Advanced Paging

- c. Go to the gene page for PITG_05224 and look at the protein feature section. Does this look like a possible motor protein?
 Hint: click on the ID for PITG_05224 in the result table to go to the gene page.
 Mouse over the glyphs in the Protein Features graphic.



6.2 Using regular expressions to find motifs in Phytophthora. Find variations of RXLR

- a. To infect plants Phytophthora utilizes effector proteins. Use a text search to find all proteins that have been identified as effectors in Phytophthora.

Text (product name, notes, etc.)

Organism

- Agaricomycetes
- Blastocladiomycetes
- Chytridiomycetes
- Eurotiomycetes
- Leotiomycetes
- Oomycetes
 - Hyaloperonospora
 - Phytophthora
 - Pythium
 - Saprolegnia
- Pneumocystidomycetes
- Pucciniomycetes
- Saccharomycetes
- Schizosaccharomycetes
- Sordariomycetes
- Tremellomycetes
- Ustilaginomycetes
- Zygomycetes

Text term (use * as wildcard)

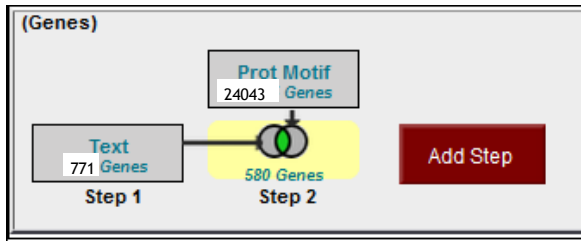
Fields Alias EC descriptions Gene ID Gene notes Gene product GO terms and definitions Metabolic pathway names and descriptions Protein domain names and descriptions Similar proteins (BLAST hits v. NRDB/FPDB) User comments

(Genes)

Text
771 Genes

Step 1

- b. RXLR is a domain motif found in some effectors to facilitate infection. Identify all occurrences of the RXLR motif in Phytophthora. You may need to refer to the RegEx guides to find the correct query; you will need to use a special character for 'X'.

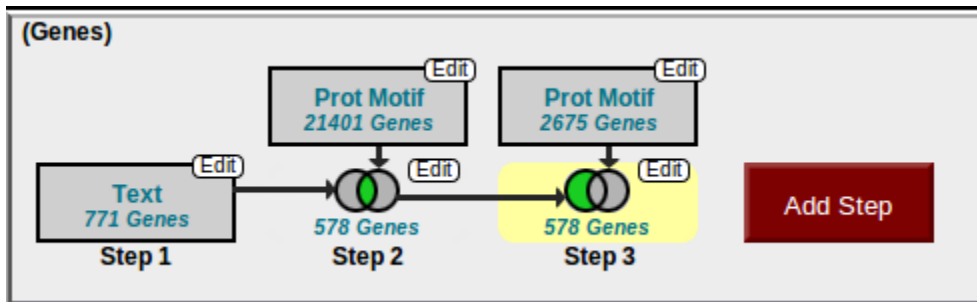


- c. Some of these were probably identified in incomplete proteins. You could use a text search to omit predicted or hypothetical proteins, but instead think of a protein motif search to only identify complete proteins containing RxLR. Protein sequences in FungiDB do not contain the stop character (*). However, bad computationally predicted proteins can have internal stops. Edit your motif search to select for proteins that start with a Methionine, do not have any *s, and contain an RXLR.

(hint: you'll need to tell the RegEx to not find * both before and after the RXLR)

You can find a single RegEx to identify the correct proteins but it will be complex. Try to break it up into multiple steps to make it easier to build.

Here it is split into two RegEx:



It only removed two genes. Why? Compare the results from B and C, where was the change?

- d. The 'X' in RXLR is a wild-card, allowing for any amino acid. Try some specific amino acids or special characters to narrow down the RXLR occurrences. Do most identified RXLRs fit into any special classification?